

# Package ‘PASWR’

February 14, 2012

**Type** Package

**Title** PROBABILITY and STATISTICS WITH R

**Version** 1.1

**LazyLoad** yes

**LazyData** yes

**Date** 2009-11-24

**Author** Alan T. Arnholt

**Maintainer** Alan T. Arnholt <arnholtat@appstate.edu>

**Depends** e1071, MASS, lattice

**Description** Data and functions for the book PROBABILITY and STATISTICS WITH R.

**License** GPL (>= 2)

**URL** <http://www1.appstate.edu/~arnholt/PASWR/>

**Repository** CRAN

**Date/Publication** 2009-11-25 08:29:13

## R topics documented:

PASWR-package . . . . .	4
Aggression . . . . .	4
Apple . . . . .	5
AptSize . . . . .	6
Baberuth . . . . .	7
Bac . . . . .	8
Battery . . . . .	9
binogen . . . . .	9
biomass . . . . .	10
Bodyfat . . . . .	11

Calculus . . . . .	12
Cars2004EU . . . . .	12
checking.plots . . . . .	13
Chips . . . . .	14
CircuitDesigns . . . . .	15
CIsim . . . . .	15
Combinations . . . . .	17
CosAma . . . . .	18
Cows . . . . .	18
Depend . . . . .	19
Drosophila . . . . .	20
EDA . . . . .	20
Engineer . . . . .	21
EPIDURAL . . . . .	22
EPIDURALf . . . . .	23
EURD . . . . .	24
fagus . . . . .	24
FCD . . . . .	25
Fertilize . . . . .	26
food . . . . .	27
Formula1 . . . . .	27
GD . . . . .	28
glucose . . . . .	29
Grades . . . . .	29
Grocery . . . . .	30
HardWater . . . . .	31
House . . . . .	32
HSwrestler . . . . .	32
Hubble . . . . .	33
InsurQuotes . . . . .	34
interval.plot . . . . .	35
janka . . . . .	35
Kinder . . . . .	36
ksdist . . . . .	37
ksLdist . . . . .	37
Leddiode . . . . .	38
LostR . . . . .	39
MilkCarton . . . . .	39
normarea . . . . .	40
nsize . . . . .	41
ntester . . . . .	42
oneway.plots . . . . .	43
Phenyl . . . . .	43
Phone . . . . .	44
Rat . . . . .	45
Ratbp . . . . .	46
Refrigerator . . . . .	46
Roacheggs . . . . .	47

Salinity	48
satfruit	49
SBIQ	50
Schizo	51
Score	51
SDS4	52
SIGN.test	53
SimDataST	55
SimDataXT	56
Soccer	56
SRS	57
StatTemps	58
Stschool	59
Sundig	59
sunflower	60
SurfaceSpain	61
Swintimes	62
Tennis	63
TestScores	63
Tire	64
TireWear	65
titanic3	65
TOE	67
Top20	67
tsum.test	68
twoway.plots	71
URLaddress	72
vit2005	72
Wait	74
Washer	74
Water	75
WCST	76
WeightGain	76
WheatSpain	77
wheatUSA2004	78
wilcoxE.test	78
Wool	80
z.test	81
zsum.test	84

---

PASWR-package

*Probability and Statistics with R*

---

### Description

Data and functions for the book *Probability and Statistics with R*

### Details

Package: PASWR  
Type: Package  
Version: 1.1  
Date: 2009-11-24  
License: GPL (>=2)

Comprehensive and engineering-oriented, *Probability and Statistics with R* provides a thorough treatment of probability and statistics, clear and accessible real-world examples, and fully detailed proofs. The text provides step-by-step explanations for numerous examples in R and S-PLUS for nearly every topic covered, including both traditional and nonparametric techniques. With a wide range of graphs to illustrate complex material as well as a solutions manual, the book also offers an accompanying website that features supporting information, including datasets, functions, and other downloadable material. It is ideal for undergraduate students and for engineers and scientists who must perform statistical analyses.

### Author(s)

Alan T. Arnholt

Maintainer: <arnholtat@appstate.edu>

### References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

---

Aggression

*TV and Behavior*

---

### Description

Data regarding aggressive behavior in relation to exposure to violent television programs used in Example 10.5

**Usage**

Aggression

**Format**

A data frame with 16 observations on the following 2 variables:

violence a numeric vector

noviolence a numeric vector

**Details**

This is data regarding aggressive behavior in relation to exposure to violent television programs from Gibbons (1997) with the following exposition:

... a group of children are matched as well as possible as regards home environment, genetic factors, intelligence, parental attitudes, and so forth, in an effort to minimize factors other than TV that might influence a tendency for aggressive behavior. In each of the resulting 16 pairs, one child is randomly selected to view the most violent shows on TV, while the other watches cartoons, situation comedies, and the like. The children are then subjected to a series of tests designed to produce an ordinal measure of their aggression factors. (pages 143-144)

**Source**

Gibbons, J. D. (1997) *Nonparametric Methods for Quantitative Analysis*. American Sciences Press.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data= Aggression,  
wilcox.test(violence, noviolence, paired=TRUE,  
alternative="greater"))
```

---

Apple

*Apple Hardness*

---

**Description**

An experiment was undertaken where seventeen recently picked (Fresh) apples were randomly selected and measured for hardness. Seventeen apples were also randomly selected from a warehouse (Warehouse) where the apples had been stored for one week. Data are used in Example 8.10.

**Usage**

Apple

**Format**

A data frame with 17 observations on the following 2 variables:

Fresh hardness rating measured in kg/meter<sup>2</sup>

Warehouse hardness rating measured in kg/meter<sup>2</sup>

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
# Figure 8.5
attach(Apple)
par(pty = "s")
Altblue <- "#A9E2FF"
Adkblue <- "#0080FF"
fresh <- qqnorm(Fresh)
old <- qqnorm(Warehouse)
plot(fresh,type="n",ylab="Sample Quantiles",xlab="Theoretical Quantiles")
qqline(Fresh, col = Altblue)
qqline(Warehouse, col = Adkblue)
points(fresh, col = Altblue, pch = 16, cex = 1.2)
points(old, col = Adkblue, pch = 17)
legend(-1.75, 9.45, c("Fresh", "Warehouse"), col = c(Altblue, Adkblue),
text.col=c("black","black"),pch=c(16,17),lty=c(1,1),bg="gray95",cex=0.75)
title("Q-Q Normal Plots")
detach(Apple)
# Trellis approach
qqmath(~c(Fresh, Warehouse), type=c("p","r"), pch=c(16,17),
cex = 1.2, col=c("#A9E2FF", "#0080FF"),
groups=rep(c("Fresh", "Warehouse"), c(length(Fresh), length(Warehouse))),
data=Apple, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
```

---

AptSize

*Apartment Size*

---

**Description**

Size of apartments in Mendebaldea, Spain and San Jorge, Spain

**Usage**

AptSize

**Format**

A data frame with 8 observations on the following 2 variables:

Mendebaldea Mendebaldea apartment size in square meters

SanJorge San Jorge apartment size in square meters

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = AptSize,
      boxplot(Mendebaldea, SanJorge) )
```

Baberuth

*George Herman Ruth*

**Description**

Baseball statistics for George Herman Ruth (The Bambino or The Sultan Of Swat)

**Usage**

Baberuth

**Format**

A data frame with 22 observations on the following 14 variables.

Year year in which the season occurred

Team team he played for Bos-A, Bos-N, or NY-A

G games played

AB at bats

R runs scored

H hits

X2B doubles

X3B triples

HR home runs

RBI runs batted in

SB stolen bases

BB base on balls or walks

BA batting average H/AB

SLG slugging percentage (total bases/at bats)

**Source**

[http://www.baseball-reference.com/about/bat\\_glossary.shtml](http://www.baseball-reference.com/about/bat_glossary.shtml)

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Baberuth,  
hist(RBI))
```

---

Bac

*Blood Alcohol Content*

---

**Description**

Two volunteers each consumed a twelve ounce beer every fifteen minutes for one hour. One hour after the fourth beer was consumed, each volunteer's blood alcohol was measured with a different breathalyzer from the same company. The numbers recorded in data frame Bac are the sorted blood alcohol content values reported with breathalyzers from company X and company Y. Data are used in Example 9.15.

**Usage**

Bac

**Format**

A data frame with 10 observations on the following 2 variables:

X blood alcohol content measured in g/L

Y blood alcohol content measured in g/L

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Bac,  
var.test(X, Y, alternative="less"))
```

---

Battery

*Lithium Batteries*

---

### Description

A manufacturer of lithium batteries has two production facilities, A and B. Fifty randomly selected batteries with an advertised life of 180 hours are selected, and tested. The lifetimes are stored in (facilityA). Fifty randomly selected batteries with an advertised life of 200 hours are selected, and tested. The lifetimes are stored in (facilityB).

### Usage

Battery

### Format

A data frame with 50 observations on the following 2 variables:

facilityA life time measured in hours

facilityB life time measured in hours

### Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
with(data = Battery,
      qqnorm(facilityA))
with(data = Battery,
      qqline(facilityA))
```

---

bino.gen

*Simulating Binomial Distribution*

---

### Description

Function that generates and displays  $m$  repeated samples of  $n$  Bernoulli trials with a given probability of success.

### Usage

```
bino.gen(samples, n, pi)
```

**Arguments**

samples            number of repeated samples to generate  
 n                    number of Bernoulli trials  
 pi                   probability of success for Bernoulli trial

**Value**

simulated.distribution            Simulated binomial distribution  
 theoretical.distribution            Theoretical binomial distribution

**Author(s)**

Alan T. Arnholt

**Examples**

```
bino.gen(1000, 20, 0.75)
```

---

biomass

*Beech Trees*

---

**Description**

Several measurements of 42 beech trees (*Fagus Sylvatica*) taken from a forest in Navarra (Spain).

**Usage**

```
biomass
```

**Format**

A data frame with 42 observations on the following 4 variables:

Dn diameter of the stem in centimeters  
 H height of the tree in meters  
 PST weight of the stem in kilograms  
 PSA aboveground weight in kilograms

**Source**

*Gobierno de Navarra and Gestion Ambiental Viveros y Repoblaciones de Navarra, 2006.* The data were obtained within the European Project FORSEE.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
plot(log(PSA)~log(Dn), data=biomass)
```

---

Bodyfat

*Body Fat Composition*

---

**Description**

Values from a study reported in the *American Journal of Clinical Nutrition* that investigated a new method for measuring body composition

**Usage**

Bodyfat

**Format**

A data frame with 18 observations on the following 3 variables:

age age in years

fat body fat composition

sex a factor with levels F for female and M for male

**Source**

Mazess, R. B., Peppler, W. W., and Gibbons, M. (1984) *Total Body Composition by Dual-Photon (153 Gd) Absorptiometry*. *American Journal of Clinical Nutrition*, **40**, **4**: 834-839.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
boxplot(fat~sex, data = Bodyfat)
```

---

Calculus

*Calculus Assessment Scores*

---

**Description**

Mathematical assessment scores for 36 students enrolled in a biostatistics course according to whether or not the students had successfully completed a calculus course prior to enrolling in the biostatistics course

**Usage**

Calculus

**Format**

A data frame with 18 observations on the following 2 variables:

No.Calculus assessment score for students with no prior calculus

Yes.Calculus assessment score for students with prior calculus

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Calculus,  
z.test(x=Yes.Calculus,y=No.Calculus,sigma.x=5,sigma.y=12)$conf )
```

---

Cars2004EU

*Cars in the European Union (2004)*

---

**Description**

The numbers of cars per 1000 inhabitants (cars), the total number of known mortal accidents (deaths), and the country population/1000 (population) for the 25 member countries of the European Union for the year 2004

**Usage**

Cars2004EU

**Format**

A data frame with 25 observations on the following 4 variables:

country a factor with levels Austria, Belgium, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, and United Kingdom

cars numbers of cars per 1000 inhabitants

deaths total number of known mortal accidents

population country population/1000

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
plot(deaths~cars, data = Cars2004EU)
```

---

checking.plots

*Checking Plots*

---

**Description**

Function that creates the three graphs discussed in Sections 11.5.1, 11.5.2, and 11.5.3 that assesses independence, normality, and constant variance

**Usage**

```
checking.plots(model, n.id = 3, COL = c("#0080FF", "#A9E2FF"))
```

**Arguments**

model	a linear model object
n.id	the number of points to identify
COL	vector of two colors

**Note**

Requires package **MASS**.

**Author(s)**

Alan T. Arnholt

**See Also**[twoway.plots](#)**Examples**

```
# Figure 11.10
mod.aov <- aov(StopDist~tire, data = Tire)
checking.plots(mod.aov)
rm(mod.aov)
```

Chips

*Silicon Chips***Description**

Two techniques of splitting chips are randomly assigned to 28 sheets so that each technique is applied to 14 sheets. The values recorded in Chips are the number of usable chips from each silicon sheet.

**Usage**

Chips

**Format**

A data frame with 14 observations on the following 2 variables:

```
techniqueI number of usable chips
techniqueII number of usable chips
```

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
par(mfrow=c(1,2))
with(data = Chips, qqnorm(techniqueI))
with(data = Chips, qqline(techniqueI))
with(data = Chips, qqnorm(techniqueII))
with(data = Chips, qqline(techniqueII))
par(mfrow=c(1,1))
# Trellis Approach
graph1 <- qqmath(~techniqueI, data = Chips, type=c("p","r"))
graph2 <- qqmath(~techniqueII, data = Chips, type=c("p","r"))
print(graph1, split=c(1, 1, 2, 1), more = TRUE)
print(graph2, split=c(2, 1, 2, 1), more = FALSE)
rm(graph1, graph2)
```

---

`CircuitDesigns`*Circuit Design Lifetime*

---

**Description**

`CircuitDesigns` contains the results from an accelerated life test used to estimate the lifetime of four different circuit designs (lifetimes in thousands of hours).

**Usage**

```
CircuitDesigns
```

**Format**

A data frame with 26 observations on the following 2 variables:

`lifetime` lifetimes in thousands of hours

`design` a factor with levels `Design1`, `Design2`, `Design3`, and `Design4`

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
bwplot(design~lifetime, data = CircuitDesigns)
```

---

`CIsim`*Confidence Interval Simulation Program*

---

**Description**

This program simulates random samples from which it constructs confidence intervals for either the population mean, the population variance, or the population proportion of successes.

**Usage**

```
CIsim(samples=100, n=30, parameter=0.5, sigma=1,  
conf.level = 0.95, type = "Mean")
```

**Arguments**

samples	the number of samples desired.
n	the size of each sample.
parameter	If constructing confidence intervals for the population mean or the population variance, parameter is the population mean (i.e., type is one of either "Mean" or "Var"). If constructing confidence intervals for the population proportion of successes, the value entered for parameter represents the population proportion of successes ( $P_i$ ), and as such, must be a number between 0 and 1.
sigma	is the population standard deviation. sigma is not required if confidence intervals are of type "Pi."
conf.level	confidence level for the graphed confidence intervals, restricted to lie between zero and one
type	character string, one of "Mean", "Var", or "Pi", or just the initial letter of each, indicating the type of confidence interval simulation to perform

**Details**

Default is to construct confidence intervals for the population mean. Simulated confidence intervals for the population variance or population proportion of successes are possible by selecting the appropriate value in the type argument.

**Value**

Graph depicts simulated confidence intervals. The number of confidence intervals that do not contain the parameter of interest are counted and reported in the commands window.

**Author(s)**

Alan T. Arnholt

**Examples**

```

CIsim(100, 30, 100, 10)
# Simulates 100 samples of size 30 from
# a normal distribution with mean 100
# and standard deviation 10. From the
# 100 simulated samples, 95% confidence
# intervals for the Mean are constructed
# and depicted in the graph.

CIsim(100, 30, 100, 10, type="Var")
# Simulates 100 samples of size 30 from
# a normal distribution with mean 100
# and standard deviation 10. From the
# 100 simulated samples, 95% confidence
# intervals for the variance are constructed
# and depicted in the graph.

CIsim(100, 50, .5, type="Pi", conf.level=.90)

```

```
# Simulates 100 samples of size 50 from
# a binomial distribution where the population
# proportion of successes is 0.5. From the
# 100 simulated samples, 90% confidence
# intervals for Pi are constructed
# and depicted in the graph.
```

---

Combinations

*Combinations*

---

### **Description**

Computes all possible combinations of n objects taken k at a time.

### **Usage**

```
Combinations(n, k)
```

### **Arguments**

n	a number
k	a number less than or equal to n

### **Value**

Returns a matrix containing the possible combinations of n objects taken k at a time.

### **See Also**

[SRS](#)

### **Examples**

```
Combinations(5,2)
# The columns in the matrix list the values of the 10 possible
# combinations of 5 things taken 2 at a time.
```

---

CosAma

*Cosmed Versus Amatek*

---

### Description

The Cosmed is a portable metabolic system. A study at Appalachian State University compared the metabolic values obtained from the Cosmed to those of a reference unit (Amatek) over a range of workloads from easy to maximal to test the validity and reliability of the Cosmed. A small portion of the results for VO<sub>2</sub> (ml/kg/min) measurements taken at a 150 watt workload are stored in CosAma.

### Usage

CosAma

### Format

A data frame with 14 observations on the following 3 variables:

subject subject number

Cosmed measured VO<sub>2</sub> with Cosmed

Amatek measured VO<sub>2</sub> with Amatek

### Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
bwplot(~(Cosmed-Amatek), data=CosAma)
```

---

Cows

*Butterfat of Cows*

---

### Description

Random samples of ten mature (five-year-old and older) and ten two-year-old cows were taken from each of five breeds. The average butterfat percentage of these 100 cows is stored in the variable butterfat with the type of cow stored in the variable breed and the age of the cow stored in the variable age.

### Usage

Cows

**Format**

A data frame with 100 observations on the following 3 variables:

butterfat average butterfat percentage

age a factor with levels 2 years old and Mature

breed a factor with levels Ayrshire, Canadian, Guernsey, Holstein-Friesian, and Jersey

**Source**

Canadian record book of purebred diary cattle.

**References**

Sokal, R. R. and Rohlf, F. J. (1994) *Biometry*. W. H. Freeman, New York, third edition.

**Examples**

```
summary(aov(butterfat~breed+age, data = Cows))
```

---

Depend

*Number of Dependent Children for 50 Families*

---

**Description**

Number of dependent children for 50 families.

**Usage**

```
data(Depend)
```

**Format**

A data frame with 50 observations on the following 4 variables.

C1 a numeric vector

number a numeric vector

Count a numeric vector

Percent a numeric vector

**Source**

Kitchens, L. J. (2003) *Basic Statistics and Data Analysis*. Duxbury

**Examples**

```
with(data = Depend, table(C1))
```

---

Drosophila

*Drosophila Melanogaster*

---

### Description

Drosophila contains per diem fecundity (number of eggs laid per female per day for the first 14 days of life) for 25 females from each of three lines of *Drosophila melanogaster*. The three lines are Nonselected (control), Resistant, and Susceptible. Data are used in Example 11.5.

### Usage

Drosophila

### Format

A data frame with 75 observations on the following 2 variables:

Fecundity number of eggs laid per female per day for the first 14 days of life

Line a factor with levels Nonselected, Resistant, and Susceptible

### Source

The original measurements are from an experiment conducted by R. R. Sokal (Sokal and Rohlf, 1994, p. 237).

### References

Sokal, R. R. and Rohlf, F. J. (1994) *Biometry*. W. H. Freeman, New York, third edition.

### Examples

```
summary(aov(Fecundity~Line, data = Drosophila))
```

---

EDA

*Exploratory Data Analysis*

---

### Description

Function that produces a histogram, density plot, boxplot, and Q-Q plot.

### Usage

```
EDA(x, trim = 0.05)
```

**Arguments**

`x` is a numeric vector where NAs and Infs are allowed but will be removed.

`trim` is a fraction (between 0 and 0.5, inclusive) of values to be trimmed from each end of the ordered data such that if `trim = 0.5`, the result is the median.

**Details**

Will not return command window information on data sets containing more than 5000 observations. It will, however, still produce graphical output for data sets containing more than 5000 observations.

**Value**

Function returns various measures of center and location. The values returned for the Quartiles are based on the definitions used in Minitab and SPSS. The boxplot is based on the Quartiles returned in the commands window. For more information on the definition of the Quartiles, type `?quantile` and read about the algorithm used by `type=6`.

**Note**

Requires package **e1071**.

**Author(s)**

Alan T. Arnholt

**Examples**

```
EDA(rnorm(100))
# Produces four graphs for the 100 randomly
# generated standard normal variates.
```

---

Engineer

*Engineer Salaries*

---

**Description**

Salaries for engineering graduates 10 years after graduation

**Usage**

Engineer

**Format**

A data frame with 51 observations on the following 2 variables:

`salary` salary 10 years after graduation in thousands of dollars  
`university` one of three different engineering universities

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
boxplot(salary~university, data = Engineer, horizontal=TRUE)
# Trellis Approach
bwplot(university~salary, data = Engineer)
```

---

 EPIDURAL

---

*Traditional Sitting Position Versus Hamstring Stretch Position*


---

**Description**

Initial results from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts (OC)

**Usage**

EPIDURAL

**Format**

A data frame with 85 observations on the following 7 variables:

Doctor a factor with levels Dr. A, Dr. B, Dr. C, and Dr. D

kg weight in kg of patient

cm height in cm of patient

Ease a factor with levels Difficult, Easy, and Impossible indicating the physician's assessment of how well bone landmarks can be felt in the patient

Treatment a factor with levels Hamstring Stretch and Traditional Sitting

OC number of obstructive contacts

Complications a factor with levels Failure - person got dizzy, Failure - too many OCs, None, Paresthesia, and Wet Tap

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```

EPIDURAL$Teasy <- factor(EPIDURAL$Ease,
  levels=c("Easy", "Difficult", "Impossible"))
X <- table(EPIDURAL$Doctor, EPIDURAL$Teasy)
X
par(mfrow=c(2,2)) # Figure 2.12
barplot(X,
  main="Barplot where Doctor is Stacked \n within Levels of Palpitation")
barplot(t(X),
  main="Barplot where Levels of Palpitation \n is Stacked within Doctor")
barplot(X, beside=TRUE,
  main="Barplot where Doctor is Grouped \n within Levels of Palpitation")
barplot(t(X), beside=TRUE,
  main="Barplot where Levels of Palpitation \n is Grouped within Doctor")
par(mfrow=c(1,1))
rm(X)

```

---

 EPIDURALf

*Traditional Sitting Position Versus Hamstring Stretch Position*


---

**Description**

Intermediate results from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts (OC)

**Usage**

```
EPIDURALf
```

**Format**

A data frame with 342 observations on the following 7 variables:

Doctor a factor with levels Dr. A, Dr. B, Dr. C, and Dr. D

kg weight in kg of patient

cm height in cm of patient

Ease a factor with levels Difficult, Easy, and Impossible indicating the physician's assessment of how well bone landmarks can be felt in the patient

Treatment a factor with levels Hamstring Stretch and Traditional Sitting

OC number of obstructive contacts

Complications a factor with levels Failure - person got dizzy, Failure - too many OCs, None, Paresthesia, and Wet Tap

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
boxplot(OC~Treatment, data = EPIDURALf)
```

---

 EURD

*European Union Research and Development*


---

**Description**

A random sample of 15 countries' research and development investments for the years 2002 and 2003 is taken and the results in millions of euros are stored in EURD.

**Usage**

```
EURD
```

**Format**

A data frame with 15 observations on the following 3 variables:

Country a factor with levels Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, France, Hungary, Latvia, Lithuania, Malta, Portugal, Romania, Slovakia, and Slovenia

RD2002 research and development investments in millions of euros for 2002

RD2003 research and development investments in millions of euros for 2003

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
qqmath(~(RD2003 - RD2002), data = EURD, type=c("p", "r"))
```

---

 fagus

*Retained Carbon in Beech Trees*


---

**Description**

The carbon retained by leaves measured in kg/ha is recorded for forty-one different plots of mountainous regions of Navarra (Spain), depending on the forest classification: areas with 90% or more beech trees (*Fagus Sylvatica*) are labeled monospecific, while areas with many species of trees are labeled multispecific.

**Usage**

```
fagus
```

**Format**

A data frame with 41 observations on the following 3 variables:

Plot plot number

carbon carbon retained by leaves measured in kg/ha

type a factor with levels monospecific and multispecific

**Source**

*Gobierno de Navarra and Gestion Ambiental Viveros y Repoblaciones de Navarra, 2006.* The data were obtained within the European Project FORSEE.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
boxplot(carbon~type, data=fagus)
```

---

FCD

*Fat Cats*

---

**Description**

In a weight loss study on obese cats, overweight cats were randomly assigned to one of three groups and boarded in a kennel. In each of the three groups, the cats' total caloric intake was strictly controlled (1 cup of generic cat food) and monitored for 10 days. The difference between the groups was that group A was given 1/4 of a cup of cat food every six hours, group B was given 1/3 a cup of cat food every eight hours, and group C was given 1/2 a cup of cat food every twelve hours. The weight of the cats at the beginning and end of the study was recorded and the difference in weights (grams) is stored in the variable `Weight` of the data frame `FCD`. Data are used in Example 11.4.

**Usage**

FCD

**Format**

A data frame with 36 observations on the following 2 variables:

Weight difference in weights (grams)

Diet a factor with levels A, B, and C

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
# Figure 11.12
FCD.aov <- aov(Weight~Diet, data = FCD)
checking.plots(FCD.aov)
rm(FCD.aov)
```

---

Fertilize

*Cross and Auto Fertilization*

---

**Description**

Plants' heights in inches obtained from two seeds, one obtained by cross fertilization and the other by auto fertilization, in two opposite but separate locations of a pot are recorded.

**Usage**

Fertilize

**Format**

A data frame with 15 observations on the following 2 variables:

cross height of plant in inches

self height of plant in inches

**Source**

Darwin, C. (1876) *The Effect of Cross and Self-Fertilization in the Vegetable Kingdom*

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Fertilize,
t.test(cross, self))
```

---

food

*Carrot Shear*

---

**Description**

Shear measured in kN on frozen carrots from four randomly selected freezers

**Usage**

food

**Format**

A data frame with 16 observations on the following 2 variables:

shear carrot shear measured in kN

freezer a factor with levels A, B, C, and D

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
summary(aov(shear~freezer, data = food))
```

---

Formula1

*Pit Stop Times*

---

**Description**

Pit stop times for two teams at 10 randomly selected Formula 1 races

**Usage**

Formula1

**Format**

A data frame with 10 observations on the following 3 variables:

Race number corresponding to a race site

Team1 pit stop times for team one

Team2 pit stop times for team two

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Formula1,  
      boxplot(Team1, Team2))
```

---

GD

*Times Until Failure*

---

**Description**

Contains time until failure in hours for a particular electronic component subjected to an accelerated stress test.

**Usage**

GD

**Format**

A data frame with 100 observations on the following variable:

attf times until failure in hours

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = GD,  
      hist(attf,prob=TRUE))  
with(data = GD,  
      lines(density(attf)))  
# Trellis Approach  
histogram(~attf, data = GD, type="density",  
          panel = function(x, ...) {  
            panel.histogram(x, ...)  
            panel.densityplot(x, col = "blue", plot.points=TRUE, lwd=2)  
          } )
```

---

glucose

*Blood Glucose Levels*

---

**Description**

Fifteen diabetic patients were randomly selected, and their blood glucose levels were measured in mg/100 ml with two different devices.

**Usage**

glucose

**Format**

A data frame with 15 observations on the following 3 variables:

Patient patient number

Old blood glucose level in mg/100 ml using old device

New blood glucose level in mg/100 ml using new device

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = glucose,  
      boxplot(Old, New))
```

---

Grades

*GPA and SAT Scores*

---

**Description**

The admissions committee of a comprehensive state university selected at random the records of 200 second semester freshmen. The results, first semester college GPA and SAT scores, are stored in the data frame Grades. Data are used in Example 12.6.

**Usage**

Grades

**Format**

A data frame with 200 observations on the following 2 variables:

sat SAT score  
gpa grade point average

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
# traditional scatterplot
plot(gpa~sat, data = Grades)
# trellis scatterplot
xyplot(gpa~sat, data = Grades, type=c("p", "smooth"))
```

---

Grocery

*Grocery Spending*

---

**Description**

The consumer expenditure survey, created by the U.S. Department of Labor, was administered to 30 households in Watauga County, North Carolina, to see how the cost of living in Watauga county with respect to total dollars spent on groceries compares with other counties. The amount of money each household spent per week on groceries is stored in the variable `groceries`. Data are used in Example 8.3.

**Usage**

Grocery

**Format**

A data frame with 30 observations on the following variable:

`groceries` total dollars spent on groceries

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Grocery,
z.test(x=groceries, sigma.x=30, conf.level=.97)$conf)
```

---

HardWater	<i>Mortality and Water Hardness</i>
-----------	-------------------------------------

---

**Description**

Mortality and drinking water hardness for 61 cities in England and Wales.

**Usage**

HardWater

**Format**

A data frame with 61 observations on the following 4 variables.

location a factor with levels North South indicating whether the town is as north as Derby

town the name of the town

mortality averaged annual mortality per 100,000 males

hardness calcium concentration (in parts per million)

**Details**

These data were collected in an investigation of environmental causes of disease. They show the annual mortality rate per 100,000 for males, averaged over the years 1958-1964, and the calcium concentration (in parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water.)

**Source**

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994) *A Handbook of Small Datasets*. Chapman and Hall/CRC, London.

**Examples**

```
plot(mortality~hardness, data=HardWater)
```

---

House

*House Prices*

---

**Description**

Random sample of house prices (in thousands of dollars) for three bedroom/two bath houses in Watauga County, NC

**Usage**

House

**Format**

A data frame with 14 observations on the following 2 variables:

Neighborhood a factor with levels Blowing Rock, Cove Creek, Green Valley, Park Valley, Parkway, and Valley Crucis

Price price of house (in thousands of dollars)

**Examples**

```
with(data = House,  
t.test(Price))
```

---

HSwrestler

*High School Wrestlers*

---

**Description**

The body fat of 78 high school wrestlers was measured using three separate techniques, and the results are stored in the data frame HSwrestler. The techniques used were hydrostatic weighing (HWFAT), skin fold measurements (SKFAT), and the Tanita body fat scale (TANFAT). Data are used in Examples 10.11, 12.11, and 12.12.

**Usage**

HSwrestler

**Format**

A data frame with 78 observations on the following 9 variables:

AGE age of wrestler in years  
HT height of wrestler in inches  
WT weight of wrestler in pounds  
ABS abdominal fat  
TRICEPS tricep fat  
SUBSCAP subscapular fat  
HWFAT hydrostatic fat  
TANFAT Tanita fat  
SKFAT skin fat

**Source**

Data provided by Dr. Alan Utter, Department of Health Leisure and Exercise Science, Appalachian State University.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
FAT <- c(HSwrestler$HWFAT, HSwrestler$TANFAT, HSwrestler$SKFAT)
GROUP <- factor(rep(c("HWFAT", "TANFAT", "SKFAT"), rep(78, 3)))
BLOCK <- factor(rep(1:78, 3))
friedman.test(FAT~GROUP|BLOCK)
```

---

Hubble

*Hubble Telescope*

---

**Description**

The Hubble Space Telescope was put into orbit on April 25, 1990. Unfortunately, on June 25, 1990, a spherical aberration was discovered in Hubble's primary mirror. To correct this, astronauts had to work in space. To prepare for the mission, two teams of astronauts practiced making repairs under simulated space conditions. Each team of astronauts went through 15 identical scenarios. The times to complete each scenario were recorded in days.

**Usage**

Hubble

**Format**

A data frame with 15 observations on the following 2 variables:

Team1 days to complete scenario

Team2 days to complete scenario

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Hubble,
      qqnorm(Team1 - Team2))
with(data = Hubble,
      qqline(Team1 - Team2))
# Trellis Approach
qqmath(~(Team1 - Team2), data = Hubble, type=c("p","r"))
```

---

InsurQuotes

*Insurance Quotes*

---

**Description**

Insurance quotes for two insurers of hazardous waste jobs

**Usage**

InsurQuotes

**Format**

A data frame with 15 observations on the following 2 variables:

companyA quotes from company A in euros

companyB quotes from company B in euros

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = InsurQuotes,
      t.test(companyA, companyB))
```

---

interval.plot	<i>Interval Plot</i>
---------------	----------------------

---

**Description**

Function to graph confidence intervals

**Usage**

```
interval.plot(ll, ul)
```

**Arguments**

ll	vector of lower confidence interval values
ul	vector of upper confidence interval values

**See Also**

[CIsim](#)

**Examples**

```
set.seed(402)
m <- 100 # Number of samples
n <- 500 # Sample size
a <- array(0,m)
ll <- array(0,m)
ul <- array(0,m)
i <- 0
while (i < m)
{ i <- i + 1
  a[i] <- mean(rnorm(n))
  ll[i] <- a[i] + qnorm(0.025)*sqrt(1/n)
  ul[i] <- a[i] + qnorm(0.975)*sqrt(1/n)
}
interval.plot(ll,ul)
```

---

janka	<i>Australian Eucalypt Hardwoods</i>
-------	--------------------------------------

---

**Description**

The dataset consists of density and hardness measurements from 36 Australian Eucalypt hardwoods.

**Usage**

```
janka
```

**Format**

A data frame with 36 observations on the following 2 variables.

Density a measure of density of the timber

Hardness the Janka hardness of the timber

**Details**

Janka Hardness is an importance rating of Australian hardwood timbers. The test measures the force required to imbed a steel ball into a piece of wood.

**Source**

Williams, E.J. (1959) *Regression Analysis*. John Wiley & Sons, New York.

**Examples**

```
with(data = janka, plot(Hardness~Density, col="blue"))
```

---

Kinder

*Kindergarten Class*

---

**Description**

The data frame Kinder contains the height in inches and weight in pounds of 20 children from a kindergarten class. Data are used in Example 12.17.

**Usage**

Kinder

**Format**

A data frame with 20 observations on the following 2 variables:

ht height in inches of child

wt weight in pounds of child

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
# Figure 12.10
with(data = Kinder,
plot(wt, ht))
# Trellis Approach
xyplot(ht~wt, data = Kinder)
```

---

ksdist                                      *Simulated Distribution of Dn (Kolmogorov-Smirnov)*

---

**Description**

Function to visualize the sampling distribution of Dn and to find simulated critical values.

**Usage**

```
ksdist(n = 10, sims = 10000, alpha = 0.05)
```

**Arguments**

n	sample size
sims	number of simulations to perform
alpha	desired alpha level

**Author(s)**

Alan T. Arnholt

**See Also**

[ksLdist](#)

**Examples**

```
ksdist(n = 10, sims = 10000, alpha = 0.05)
```

---

ksLdist                                      *Simulated Lilliefors' Test of Normality Values*

---

**Description**

Function to visualize the sampling distribution of Dn for simple and composite hypotheses

**Usage**

```
ksLdist(n = 10, sims = 1000, alpha = 0.05)
```

**Arguments**

n	sample size
sims	number of simulations
alpha	desired alpha level

**Author(s)**

Alan T. Arnholt

**See Also**

[ksdist](#)

**Examples**

```
# Figure 10.15
ksLdist(sims=10000,n=10)
```

---

Leddiode

*LED Diodes*

---

**Description**

The diameter in millimeters for a random sample of 15 diodes from each of the two suppliers is stored in the data frame `Leddiode`.

**Usage**

```
Leddiode
```

**Format**

A data frame with 15 observations on the following 2 variables:

`supplierA` diameter in millimeters of diodes from supplier A

`supplierB` diameter in millimeters of diodes from supplier B

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Leddiode,
      boxplot(supplierA, supplierB, col=c("red", "blue")))
```

---

LostR

*Lost Revenue Due to Worker Illness*

---

**Description**

Data set containing the lost revenue in dollars/day and number of workers absent due to illness for a metallurgic company

**Usage**

LostR

**Format**

A data frame with 25 observations on the following 2 variables:

NumberSick number of absent workers due to illness

LostRevenue lost revenue in dollars

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
xyplot(LostRevenue~NumberSick, data = LostR, type=c("p","r"))
```

---

MilkCarton

*Milk Carton Drying Times*

---

**Description**

A plastics manufacturer makes two sizes of milk containers: half gallon and gallon sizes. The time required for each size to dry is recorded in seconds in the data frame MilkCarton.

**Usage**

MilkCarton

**Format**

A data frame with 40 observations on the following 2 variables:

Hgallon drying time in seconds for half gallon containers

Wgallon drying time in seconds for whole gallon containers

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = MilkCarton,  
      boxplot(Hgallon, Wgallon))
```

---

normarea

*Normal Area*

---

**Description**

Function that computes and draws the area between two user specified values in a user specified normal distribution with a given mean and standard deviation

**Usage**

```
normarea(lower = -Inf, upper = Inf, m, sig)
```

**Arguments**

lower	the lower value
upper	the upper value
m	the mean for the population
sig	the standard deviation of the population

**Author(s)**

Alan T. Arnholt

**Examples**

```
normarea(70, 130, 100, 15)  
# Finds and P(70 < X < 130) given X is N(100,15).
```

---

nsize	<i>Required Sample Size</i>
-------	-----------------------------

---

**Description**

Function to determine required sample size to be within a given margin of error

**Usage**

```
nsize(b, sigma = NULL, p = 0.5, conf.level = 0.95, type = "mu")
```

**Arguments**

b	the desired bound
sigma	population standard deviation; not required if using type "pi"
p	estimate for the population proportion of successes; not required if using type "mu"
conf.level	confidence level for the problem, restricted to lie between zero and one
type	character string, one of "mu" or "pi", or just the initial letter of each, indicating the appropriate parameter; default value is "mu"

**Details**

Answer is based on a normal approximation when using type "pi".

**Value**

Returns required sample size.

**Author(s)**

Alan T. Arnholt

**Examples**

```
nsize(b=0.015, p=0.5, conf.level=.95,type="pi")
# Returns the required sample size (n) to estimate the population
# proportion of successes with a 0.95 confidence interval
# so that the margin of error is no more than 0.015 when the
# estimate of the population propotion of successes is 0.5.
# This is Example 8.25 part b. from PASWR.
```

```
nsize(b=0.02, sigma=0.1,conf.level=.95,type="mu")
# Returns the required sample size (n) to estimate the population
# mean with a 0.95 confidence interval so that the margin
# of error is no more than 0.02. This is Example 8.4 from PASWR.
```

---

ntester

*Normality Tester*

---

### Description

Q-Q plots of randomly generated normal data of the same sample size as the tested data are generated and plotted on the perimeter of the graph while a Q-Q plot of the actual data is depicted in the center of the graph.

### Usage

```
ntester(actual.data)
```

### Arguments

`actual.data` is a numeric vector. Missing and infinite values are allowed, but are ignored in the calculation. The length of `actual.data` must be less than 5000 after dropping nonfinite values.

### Details

Q-Q plots of randomly generated normal data of the same size as the tested data are generated and plotted on the perimeter of the graph sheet while a Q-Q plot of the actual data is depicted in the center of the graph. The p-values are calculated based on the Shapiro-Wilk W-statistic. Function will only work on numeric vectors containing less than or equal to 5000 observations.

### Author(s)

Alan T. Arnholt

### References

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591-611.

### Examples

```
ntester(rexp(50,1))
# Q-Q plot of random exponential data in center plot
# surrounded by 8 Q-Q plots of randomly generated
# standard normal data of size 50.
```

---

`oneway.plots`*Exploratory Graphs for Single Factor Designs*

---

**Description**

Function to create dotplots, boxplots, and design plot (means) for single factor designs

**Usage**

```
oneway.plots(Y, fac1, COL = c("#A9E2FF", "#0080FF"))
```

**Arguments**

Y	response variable for a single factor design
fac1	predictor variable
COL	a vector with two colors

**Author(s)**

Alan T. Arnholt

**See Also**

[twoway.plots](#)

**Examples**

```
with(data = Tire,  
oneway.plots(StopDist, tire))
```

---

`Phenyl`*Phenylketonuria*

---

**Description**

The data frame `Phenyl` records the level of Q10 at four different times for 46 patients diagnosed with phenylketonuria. The variable `Q10.1` contains the level of Q10 measured in micromoles for the 46 patients. `Q10.2`, `Q10.3`, and `Q10.4` are the values recorded at later times respectively for the 46 patients.

**Usage**

`Phenyl`

**Format**

A data frame with 46 observations on the following 4 variables.

Q10.1 level of Q10 at time 1 in micromoles

Q10.2 level of Q10 at time 2 in micromoles

Q10.3 level of Q10 at time 3 in micromoles

Q10.4 level of Q10 at time 4 in micromoles

**Details**

Phenylketonuria (PKU) is a genetic disorder that is characterized by an inability of the body to utilize the essential amino acid, phenylalanine. Research suggests patients with phenylketonuria have deficiencies in coenzyme Q10.

**Source**

Artuch, R., *et. al.* (2004) "Study of Antioxidant Status in Phenylketonuric Patients." *Clinical Biochemistry*, **37**: 198-203.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Phenyl,  
t.test(Q10.1,conf.level=.99))
```

---

Phone

*Telephone Call Times*

---

**Description**

Phone contains times in minutes of long distance telephone calls during a one month period for a small business. Data are used in Example 10.1.

**Usage**

Phone

**Format**

A data frame with 23 observations on the following variable:

call.time time spent on long distance calls in minutes

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Phone,  
SIGN.test(call.time,md=2.1))
```

---

Rat

*Rat Survival Time*

---

**Description**

The survival time in weeks of 20 male rats exposed to high levels of radiation.

**Usage**

Rat

**Format**

A data frame with 20 observations on the following variable:

survival.time number of weeks survived

**Source**

Lawless, J. (1982) *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Rat,  
EDA(survival.time))
```

---

 Ratbp

*Rat Blood Pressure*


---

### Description

Twelve rats were chosen, and a drug was administered to six rats, the treatment group, chosen at random. The other six rats, the control group, received a placebo. The drops in blood pressure (mmHg) for the treatment group (with probability distribution F) and the control group (with probability distribution G) are stored in the variables Treat and Cont, respectively. Data are used in Example 10.18.

### Usage

Ratbp

### Format

A data frame with 6 observations on the following 2 variables:

Treat drops in blood pressure in mmHg for treatment group

Cont drops in blood pressure in mmHg for control group

### Source

The data is originally from Ott and Mendenhall (1985, problem 8.17).

### References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
with(data = Ratbp,
      boxplot(Treat, Cont))
```

---

 Refrigerator

*Refrigerator Energy Consumption*


---

### Description

Thirty 18 cubic feet refrigerators were randomly selected from a company's warehouse. The first fifteen had their motors modified while the last fifteen were left intact. The energy consumption (kilowatts) for a 24 hour period for each refrigerator was recorded and stored in the data frame Refrigerator. The refrigerators with the design modification are stored in the variable modelA, and those without the design modification are stored in the variable modelB.

**Usage**

Refrigerator

**Format**

A data frame with 30 observations on the following 2 variables.

modelA energy consumption in kilowatts for a 24 hour period

modelB energy consumption in kilowatts for a 24 hour period

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Refrigerator,  
      boxplot(modelA, modelB))
```

---

Roacheggs

*Oriental Cockroaches*

---

**Description**

A laboratory is interested in testing a new child friendly pesticide on *Blatta orientalis* (oriental cockroaches). Scientists apply the new pesticide to 81 randomly selected *Blatta orientalis* oothecae (eggs). The results from the experiment are stored in the data frame Roacheggs in the variable eggs. A zero in the variable eggs indicates that nothing hatched from the egg while a 1 indicates the birth of a cockroach. Data is used in Example 7.16.

**Usage**

```
data(Roacheggs)
```

**Format**

A data frame with 81 observations on the following variable:

eggs numeric vector where a 0 indicates nothing hatched while a 1 indicates the birth of a cockroach.

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
p <- seq(0.1, 0.9, 0.001)
negloglike <- function(p){
  -(sum(Roacheggs$eggs)*log(p) + sum(1-Roacheggs$eggs)*log(1-p))
}
nlm(negloglike, .2)
rm(negloglike)
```

---

Salinity

*Surface Water Salinity*

---

**Description**

Surface-water salinity measurements were taken in a bottom-sampling project in Whitewater Bay, Florida. These data are stored in the data frame `Salinity`.

**Usage**

```
Salinity
```

**Format**

A data frame with 48 observations on the following variable:

```
salinity surface-water salinity measurements
```

**Source**

Davis, J. (1986) *Statistics and Data Analysis in Geology*. John Wiley, New York.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Salinity,
     EDA(salinity))
```

---

 satfruit

 Fruit Trees
 

---

### Description

To estimate the total surface occupied by fruit trees in 3 small areas (R63, R67, and R68) of Navarra (Spain) in 2001, a sample of 47 square segments has been taken. The experimental units are square segments or quadrats of 4 hectares, obtained by random sampling after overlaying a square grid on the study domain. Data are used in Case Study: Fruit Trees, Chapter 12.

### Usage

satfruit

### Format

A data frame with 47 observations on the following 17 variables:

QUADRAT number of the sampled segment or quadrat

SArea the small area, a factor with levels R63, R67, and R68

WH area classified as wheat in sampled segment

BA area classified as barley in sampled segment

NAR area classified as non arable in sampled segment

COR area classified as corn in sampled segment

SF area classified as sunflower in sampled segment

VI area classified as vineyard in sampled segment

PS area classified as grass in sampled segment

ES area classified as asparagus in sampled segment

AF area classified as lucerne in sampled segment

CO area classified as rape (*Brassica Napus*) in sampled segment

AR area classified as rice in sampled segment

AL area classified as almonds in sampled segment

OL area classified as olives in sampled segment

FR area classified as fruit trees in sampled segment

OBS the observed area of fruit trees in sampled segment

### Source

Militino, A. F., *et. al.* (2006) "Using Small Area Models to Estimate the Total Area Occupied by Olive Trees." *Journal of Agricultural, Biological and Environmental Statistics*, **11**: 450-461.

## References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

## Examples

```
with(data = satfruit,
pairs(satfruit[,15:17]))
# Trellis Approach
splom(~data.frame(satfruit[,15:17]), data = satfruit)
```

---

SBIQ

*County IQ*

---

## Description

A school psychologist administered the Stanford-Binet intelligence quotient (IQ) test in two counties. Forty randomly selected, gifted and talented students were selected from each county. The Stanford-Binet IQ test is said to follow a normal distribution with a mean of 100 and standard deviation of 16.

## Usage

SBIQ

## Format

A data frame with 40 observations on the following 2 variables:

County1 IQ scores for county one

County2 IQ scores for county two

## Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

## Examples

```
with(data = SBIQ, qqnorm(County1))
with(data = SBIQ, qqline(County1))
# Trellis Approach
qqmath(~County1, data = SBIQ, type=c("p","r"))
```

---

Schizo	<i>Dopamine Activity</i>
--------	--------------------------

---

**Description**

Twenty-five patients with schizophrenia were classified as psychotic or nonpsychotic after being treated with an antipsychotic drug. Samples of cerebral fluid were taken from each patient and assayed for dopamine b-hydroxylase (DBH) activity. The dopamine measurements for the two groups are in nmol/(ml)(h)/(mg) of protein.

**Usage**

Schizo

**Format**

A data frame with 15 observations on the following 2 variables:

nonpsychotic dopamine activity level for patients classified nonpsychotic  
 psychotic dopamine activity level for patients classified psychotic

**Source**

Sternberg, D. E., Van Kammen, D. P., and Bunney, W. E. (1982) "Schizophrenia: Dopamine b-Hydroxylase Activity and Treatment Response." *Science*, **216**: 1423-1425.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data=Schizo, boxplot(nonpsychotic, psychotic,
names=c("nonpsychotic", "psychotic"), col=c("green", "red")))
```

---

Score	<i>Standardized Test Scores</i>
-------	---------------------------------

---

**Description**

Standardized test scores from a random sample of twenty college freshmen.

**Usage**

Score

**Format**

A data frame with 20 observations on the following variable:

scores standardized test score

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
qqmath(~scores, data = Score, type=c("p", "r"))
```

---

SDS4

*M1 Motorspeedway Times*

---

**Description**

The times recorded are those for 41 successive vehicles travelling northwards along the M1 motorway in England when passing a fixed point near Junction 13 in Bedfordshire on Saturday, March 23, 1985. After subtracting the times, the following 40 interarrival times reported to the nearest second are stored in SDS4 under the variable Times. Data are used in Example 10.17.

**Usage**

SDS4

**Format**

A data frame with 40 observations on the following variable:

Times interarrival times to the nearest second

**Source**

Hand, D. J., *et. al.* (1994) *A Handbook of Small Data Sets*. Chapman & Hall, London.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = SDS4, hist(Times))
```

---

SIGN.test	<i>Sign Test</i>
-----------	------------------

---

**Description**

This function will test a hypothesis based on the sign test and reports linearly interpolated confidence intervals for one sample problems.

**Usage**

```
SIGN.test(x, y = NULL, md = 0, alternative = "two.sided", conf.level = 0.95)
```

**Arguments**

x	numeric vector; NAs and Infs are allowed but will be removed.
y	optional numeric vector; NAs and Infs are allowed but will be removed.
md	a single number representing the value of the population median specified by the null hypothesis
alternative	is a character string, one of "greater", "less", or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, alternative refers to the true median of the parent population in relation to the hypothesized value of the median.
conf.level	confidence level for the returned confidence interval, restricted to lie between zero and one

**Details**

Computes a "Dependent-samples Sign-Test" if both x and y are provided. If only x is provided, computes the "Sign-Test".

**Value**

A list of class `htest`, containing the following components:

statistic	the S-statistic (the number of positive differences between the data and the hypothesized median), with names attribute "S".
p.value	the p-value for the test
conf.int	is a confidence interval (vector of length 2) for the true median based on linear interpolation. The confidence level is recorded in the attribute <code>conf.level</code> . When the alternative is not "two.sided", the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values k for which one would not reject the null hypothesis that the true mean or difference in means is k. Here infinity will be represented by <code>Inf</code> .
estimate	is a vector of length 1, giving the sample median; this estimates the corresponding population parameter. Component <code>estimate</code> has a names attribute describing its elements.

<code>null.value</code>	is the value of the median specified by the null hypothesis. This equals the input argument <code>md</code> . Component <code>null.value</code> has a <code>names</code> attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : "greater", "less", or "two.sided"
<code>data.name</code>	a character string (vector of length 1) containing the actual name of the input vector <code>x</code>

### Null Hypothesis

For the one-sample sign-test, the null hypothesis is that the median of the population from which `x` is drawn is `md`. For the two-sample dependent case, the null hypothesis is that the median for the differences of the populations from which `x` and `y` are drawn is `md`. The alternative hypothesis indicates the direction of divergence of the population median for `x` from `md` (i.e., "greater", "less", "two.sided".)

### Assumptions

The median test assumes the parent population is continuous.

### Confidence Interval

A linear interpolation is returned for the related confidence interval (returned component `conf.int`) which can be obtained by interpolating between the possible achieved confidence levels closest to the desired level. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when `alternative` is not "two.sided"; infinity will be represented by `Inf`.

### Note

The reported confidence interval is based on linear interpolation. The lower and upper confidence levels are exact.

### Author(s)

Alan T. Arnholt

### References

- Gibbons, J.D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference*. Marcel Dekker Inc., New York.
- Kitchens, L.J.(2003). *Basic Statistics and Data Analysis*. Duxbury.
- Conover, W. J. (1980). *Practical Nonparametric Statistics, 2nd ed.* Wiley, New York.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden and Day, San Francisco.

### See Also

[z.test](#), [zsum.test](#), [tsum.test](#)

**Examples**

```
with(data = Phone,
SIGN.test(call.time,md=2.1))
# Example 10.1 from PASWR.
# Computes two-sided sign-test for the null hypothesis
# that the population median is 2.1. The alternative
# hypothesis is that the median is not 2.1. An interpolated
# upper 95% upper bound for the population median will be computed.
```

---

SimDataST

*Simulated Data (Predictors)*

---

**Description**

Simulated data for five variables. Data are used with Example 12.21.

**Usage**

SimDataST

**Format**

A data frame with 200 observations on the following 5 variables:

Y1 a numeric vector

Y2 a numeric vector

x1 a numeric vector

x2 a numeric vector

x3 a numeric vector

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
xypplot(Y1~x1, data = SimDataST, type=c("p","smooth"))
```

---

SimDataXT

*Simulated Data (Logarithms)*

---

**Description**

Simulated data for four variables. Data are used with Example 12.18.

**Usage**

```
data(SimDataXT)
```

**Format**

A data frame with 200 observations on the following 4 variables:

Y a numeric vector

x1 a numeric vector

x2 a numeric vector

x3 a numeric vector

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
xyplot(Y~x1, data = SimDataXT, type=c("p", "smooth"))
```

---

Soccer

*World Cup Soccer*

---

**Description**

Soccer contains how many goals were scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002. Data are used in Example 4.4.

**Usage**

```
Soccer
```

**Format**

A data frame with 575 observations on the following 3 variables:

CGT cumulative goal time in minutes

Game game in which goals were scored

Goals number of goals scored in regulation period

**Details**

The World Cup is played once every four years. National teams from all over the world compete. In 2002 and in 1998, thirty-six teams were invited; whereas, in 1994 and in 1990, only 24 teams participated. The data frame Soccer contains three columns: CGT, Game, and Goals. All of the information contained in Soccer is indirectly available from the FIFA World Cup website, located at <http://fifaworldcup.yahoo.com/>.

**Source**

Chu, S. (2003) "Using Soccer Goals to Motivate the Poisson Process." *INFORMS Transaction on Education*, **3**, **2**: 62-68.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Soccer,  
      table(Goals))
```

---

SRS

*Simple Random Sampling*

---

**Description**

Computes all possible samples from a given population using simple random sampling.

**Usage**

```
SRS(POPvalues, n)
```

**Arguments**

POPvalues      vector containing the population values

n                the sample size

**Value**

Returns a matrix containing the possible simple random samples of size  $n$  taken from a population POPvalues.

**Author(s)**

Alan T. Arnholt

**See Also**

[Combinations](#)

**Examples**

```
SRS(c(5,8,3),2)
# The rows in the matrix list the values for the 3 possible
# simple random samples of size 2 from the population of 5,8, and 3.
```

---

StatTemps

*Student Temperature*

---

**Description**

In a study conducted at Appalachian State University, students used digital oral thermometers to record their temperatures each day they came to class. A randomly selected day of student temperatures is provided in StatTemps. Information is also provided with regard to subject gender and the hour of the day when the students' temperatures were measured.

**Usage**

```
StatTemps
```

**Format**

A data frame with 34 observations on the following 3 variables:

temperature temperature in fahrenheit

gender a factor with levels Female and Male

class a factor with levels 8 a.m. and 9 a.m.

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
bwplot(gender~temperature, data = StatTemps)
```

---

Stschool	<i>School Satisfaction</i>
----------	----------------------------

---

**Description**

A questionnaire is randomly administered to 11 students from State School X and to 15 students from State School Y (the results have been ordered and stored in the data frame `Stschool`). Data are used in Example 9.11.

**Usage**

`Stschool`

**Format**

A data frame with 26 observations on the following 4 variables:

X satisfaction score

Y satisfaction score

Satisfaction combined satisfaction scores

School a factor with levels X and Y

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Stschool,  
t.test(X, Y, var.equal=TRUE))
```

---

Sundig	<i>Workstation Comparison</i>
--------	-------------------------------

---

**Description**

To compare the speed differences between two different brands of workstations (Sun and Digital), the times each brand took to complete complex simulations was recorded. Five complex simulations were selected, and the five selected simulations were run on both workstations. The resulting times in minutes for the five simulations are stored in data frame `Sundig`.

**Usage**

`Sundig`

**Format**

A data frame with 5 observations on the following 3 variables:

SUN time in seconds for a Sun workstation to complete a simulation

DIGITAL time in seconds for a Digital workstation to complete a simulation

d difference between Sun and Digital

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Sundig,
      t.test(SUN, DIGITAL, paired=TRUE)$conf)
```

---

sunflower

*Sunflower Defoliation*

---

**Description**

Seventy-two field trials were conducted by applying four defoliation treatments (non defoliated control, 33%, 66%, and 100%) at different growth stages (*stage*) ranging from pre-flowering (1) to physiological maturity (5) in four different locations of Navarra, Spain: Carcastillo (1), Melida (2), Murillo (3), and Unciti (4). There are two response variables: yield in kg/ha of the sunflower and numseed, the number of seeds per sunflower head. Data are stored in the data frame sunflower. Data used in Case Study: Sunflower defoliation from Chapter 11.

**Usage**

```
sunflower
```

**Format**

A data frame with 72 observations on the following 5 variables:

*location* a factor with levels A, B, C, and D for locations Carcastillo, Melida, Murillo, and Unciti respectively

*stage* a factor with levels stage1, stage2, stage3, stage4, and stage5

*defoli* a factor with levels control, treat1, treat2, and treat3

*yield* sunflower yield in kg/ha

*numseed* number of seeds per sunflower head

**Source**

Muro, J., *et al.* (2001) "Defoliation Effects on Sunflower Yield Reduction." *Agronomy Journal*, **93**: 634-637.

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
summary(aov(yield~stage+defoli+stage:defoli, data = sunflower))
```

---

SurfaceSpain

*Surface Area for Spanish Communities*

---

**Description**

Surface area (km<sup>2</sup>) for seventeen autonomous Spanish communities.

**Usage**

SurfaceSpain

**Format**

A data frame with 17 observations on the following 2 variables:

community a factor with levels Andaluca, Aragon, Asturias, Baleares, C.Valenciana, Canarias, Cantabria, Castilla-La Mancha, Castilla-Leon, Cataluna, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarra, and P.Vasco

surface surface area in km<sup>2</sup>

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = SurfaceSpain,  
barplot(surface, names.arg=community, las=2))  
# Trellis Approach  
barchart(community~surface, data=SurfaceSpain)
```

Swimtimes

*Swim Times*

---

**Description**

Swimmers' improvements in seconds for two diets are stored in the data frame `Swimtimes`. The values in `highfat` represent the time improvement in seconds for swimmers on a high fat diet, and the values in `lowfat` represent the time improvement in seconds for swimmers on a low fat diet. Data are used in Example 10.9.

**Usage**`Swimtimes`**Format**

A data frame with 14 observations on the following 2 variables:

`lowfat` time improvement in seconds

`highfat` time improvement in seconds

**Details**

Times for the thirty-two swimmers for the 200 yard individual medley were taken right after the swimmers' conference meet. The swimmers were randomly assigned to follow one of the diets. The group on diet 1 followed a low fat diet the entire year but lost two swimmers along the way. The group on diet 2 followed the high fat diet the entire year and also lost two swimmers.

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Swimtimes,  
      wilcox.test(highfat, lowfat))
```

---

Tennis                      *Speed Detector*

---

**Description**

The Yonalasee tennis club has two systems to measure the speed of a tennis ball. The local tennis pro suspects one system (Speed1) consistently records faster speeds. To test her suspicions, she sets up both systems and records the speeds of 12 serves (three serves from each side of the court). The values are stored in the data frame Tennis in the variables Speed1 and Speed2. The recorded speeds are in kilometers per hour.

**Usage**

Tennis

**Format**

A data frame with 12 observations on the following 2 variables:

Speed1 speed in kilometers per hour

Speed2 speed in kilometers per hour

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Tennis,  
      boxplot(Speed1, Speed2))
```

---

TestScores                      *Statistics Grades*

---

**Description**

Test grades of 29 students taking a basic statistics course

**Usage**

TestScores

**Format**

A data frame with 29 observations on the following variable:

grade test score

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = TestScores,  
      EDA(grade))
```

---

Tire

*Tire Stopping Distances*

---

**Description**

The data frame `Tire` has the stopping distances measured to the nearest foot for a standard sized car to come to a complete stop from a speed of sixty miles per hour. There are six measurements of the stopping distance for four different tread patterns labeled A, B, C, and D. The same driver and car were used for all twenty-four measurements. Data are used in Example 11.1 and 11.2.

**Usage**

```
Tire
```

**Format**

A data frame with 24 observations on the following 2 variables:

`StopDist` stopping distance measured to the nearest foot

`tire` a factor with levels A, B, C, and D

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
summary(aov(StopDist~tire, data = Tire))
```

---

TireWear

*Tire Wear*

---

### Description

The data frame `TireWear` contains measurements for the amount of tread loss after 10,000 miles of driving in thousandths of an inch. Data are used in Example 11.8.

### Usage

`TireWear`

### Format

A data frame with 16 observations on the following 3 variables:

`Wear` tread loss measured in thousandths of an inch

`Treat` a factor with levels A, B, C, and D

`Block` a factor with levels Car1, Car2, Car3, and Car4

### Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
par(mfrow=c(1,2), cex=.8)
with(data = TireWear,
  interaction.plot(Treat,Block,Wear, type="b", legend=FALSE))
with(data = TireWear,
  interaction.plot(Block,Treat,Wear, type="b", legend=FALSE))
par(mfrow=c(1,1), cex=1)
```

---

titanic3

*Titanic Survival Status*

---

### Description

The `titanic3` data frame describes the survival status of individual passengers on the Titanic. The `titanic3` data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

### Usage

`titanic3`

**Format**

A data frame with 1309 observations on the following 14 variables:

pclass a factor with levels 1st, 2nd, and 3rd  
survived Survival (0 = No; 1 = Yes)  
name Name  
sex a factor with levels female and male  
age age in years  
sibsp Number of Siblings/Spouses Aboard  
parch Number of Parents/Children Aboard  
ticket Ticket Number  
fare Passenger Fare  
cabin Cabin  
embarked a factor with levels Cherbourg, Queenstown, and Southampton  
boat Lifeboat  
body Body Identification Number  
home.dest Home/Destination

**Details**

Thomas Cason of UVa has greatly updated and improved the `titanic` data frame using the *Encyclopedia Titanica* and created a new dataset called `titanic3`. This dataset reflects the state of data available as of August 2, 1999. Some duplicate passengers have been dropped, many errors have been corrected, many missing ages have been filled in, and new variables have been created.

**Source**

<http://biostat.mc.vanderbilt.edu/twiki/pub/Main/DataSets/titanic.html>

**References**

Harrell, F. E. (2001) *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

**Examples**

```
with(titanic3,  
table(pclass, sex))
```

---

TOE	<i>Nuclear Energy</i>
-----	-----------------------

---

**Description**

Nuclear energy (in TOE, tons of oil equivalent) produced in 12 randomly selected European countries during 2003

**Usage**

TOE

**Format**

A data frame with 12 observations on the following variable:

energy nuclear energy measured in tons of oil equivalent

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(TOE,  
plot(density(energy)))
```

---

Top20	<i>Tennis Income</i>
-------	----------------------

---

**Description**

Top20 contains data (in millions of dollars) corresponding to the earnings of 15 randomly selected tennis players whose earnings fall somewhere in positions 20 through 100 of ranked earnings.

**Usage**

Top20

**Format**

A data frame with 15 observations on the following variable:

income yearly income in millions of dollars

**Source**

<http://www.atptennis.com/en/>

**References**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Top20,
     EDA(income))
```

---

tsum.test

*Summarized t-test*

---

**Description**

Performs a one-sample, two-sample, or a Welch modified two-sample t-test based on user supplied summary information. Output is identical to that produced with `t.test`.

**Usage**

```
tsum.test(mean.x, s.x = NULL, n.x = NULL, mean.y = NULL, s.y = NULL,
          n.y = NULL, alternative = "two.sided", mu = 0, var.equal = FALSE,
          conf.level = 0.95)
```

**Arguments**

<code>mean.x</code>	a single number representing the sample mean of <code>x</code>
<code>s.x</code>	a single number representing the sample standard deviation for <code>x</code>
<code>n.x</code>	a single number representing the sample size for <code>x</code>
<code>mean.y</code>	a single number representing the sample mean of <code>y</code>
<code>s.y</code>	a single number representing the sample standard deviation for <code>y</code>
<code>n.y</code>	a single number representing the sample size for <code>y</code>
<code>alternative</code>	is a character string, one of "greater", "less" or "two.sided", or just the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, <code>alternative</code> refers to the true mean of the parent population in relation to the hypothesized value <code>mu</code> . For the standard two-sample tests, <code>alternative</code> refers to the difference between the true population mean for <code>x</code> and that for <code>y</code> , in relation to <code>mu</code> . For the one-sample and paired t-tests, <code>alternative</code> refers to the true mean of the parent population in relation to the hypothesized value <code>mu</code> . For the standard and Welch modified two-sample t-tests, <code>alternative</code> refers to the difference between the true population mean for <code>x</code> and that for <code>y</code> , in relation to <code>mu</code> . For the one-sample t-tests, <code>alternative</code> refers to

the true mean of the parent population in relation to the hypothesized value  $\mu$ . For the standard and Welch modified two-sample t-tests, `alternative` refers to the difference between the true population mean for  $x$  and that for  $y$ , in relation to  $\mu$ .

<code>mu</code>	is a single number representing the value of the mean or difference in means specified by the null hypothesis.
<code>var.equal</code>	logical flag: if TRUE, the variances of the parent populations of $x$ and $y$ are assumed equal. Argument <code>var.equal</code> should be supplied only for the two-sample tests.
<code>conf.level</code>	is the confidence level for the returned confidence interval; it must lie between zero and one.

### Details

If `y` is NULL, a one-sample t-test is carried out with  $x$ . If `y` is not NULL, either a standard or Welch modified two-sample t-test is performed, depending on whether `var.equal` is TRUE or FALSE.

### Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the t-statistic, with names attribute <code>"t"</code>
<code>parameters</code>	is the degrees of freedom of the t-distribution associated with <code>statistic</code> . Component <code>parameters</code> has names attribute <code>"df"</code> .
<code>p.value</code>	the p-value for the test.
<code>conf.int</code>	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute <code>conf.level</code> . When <code>alternative</code> is not <code>"two.sided"</code> , the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values $k$ for which one would not reject the null hypothesis that the true mean or difference in means is $k$ . Here infinity will be represented by <code>Inf</code> .
<code>estimate</code>	vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	the value of the mean or difference in means specified by the null hypothesis. This equals the input argument <code>mu</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : <code>"greater"</code> , <code>"less"</code> or <code>"two.sided"</code> .
<code>data.name</code>	a character string (vector of length 1) containing the names $x$ and $y$ for the two summarized samples.

### Null Hypothesis

For the one-sample t-test, the null hypothesis is that the mean of the population from which  $x$  is drawn is  $\mu$ . For the standard and Welch modified two-sample t-tests, the null hypothesis is that the population mean for  $x$  less that for  $y$  is  $\mu$ .

The alternative hypothesis in each case indicates the direction of divergence of the population mean for  $x$  (or difference of means for  $x$  and  $y$ ) from  $\mu$  (i.e., "greater", "less", or "two.sided").

### Test Assumptions

The assumption of equal population variances is central to the standard two-sample t-test. This test can be misleading when population variances are not equal, as the null distribution of the test statistic is no longer a t-distribution. If the assumption of equal variances is doubtful with respect to a particular dataset, the Welch modification of the t-test should be used.

The t-test and the associated confidence interval are quite robust with respect to level toward heavy-tailed non-Gaussian distributions (e.g., data with outliers). However, the t-test is non-robust with respect to power, and the confidence interval is non-robust with respect to average length, toward these same types of distributions.

### Confidence Intervals

For each of the above tests, an expression for the related confidence interval (returned component `conf.int`) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when alternative is not "two.sided" ; infinity will be represented by `Inf`.

### Author(s)

Alan T. Arnholt

### References

- Kitchens, L.J. (2003). *Basic Statistics and Data Analysis*. Duxbury.
- Hogg, R. V. and Craig, A. T. (1970). *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.
- Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.
- Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

### See Also

[z.test](#), [zsum.test](#)

### Examples

```
round(tsum.test(mean.x=53/15, mean.y=77/11, s.x=sqrt((222-15*(53/15)^2)/14),
s.y=sqrt((560-11*(77/11)^2)/10), n.x=15, n.y=11, var.equal= TRUE)$conf, 2)
# Example 8.13 from PASWR
tsum.test(mean.x=4, s.x=2.89, n.x=25, mu=2.5)
# Example 9.8 from PASWR
```

**Description**

Function creates side-by-side boxplots for each factor, a design plot (means), and an interaction plot.

**Usage**

```
tway.plots(Y, fac1, fac2, COL = c("#A9E2FF", "#0080FF"))
```

**Arguments**

Y	response variable
fac1	factor one
fac2	factor two
COL	a vector with two colors

**Author(s)**

Alan T. Arnholt

**See Also**

[oneway.plots](#)

**Examples**

```
# Figure 11.25
Microamps <- c(280, 290, 285, 300, 310, 295, 270, 285, 290, 230,
235, 240, 260, 240, 235, 220, 225, 230)
Glass <- factor(c(rep("Glass I", 9), rep("Glass II", 9)))
Phosphor <- factor((rep(c(rep("Phosphor A", 3), rep("Phosphor B", 3),
rep("Phosphor C", 3)), 2)))
tway.plots(Microamps, Glass, Phosphor)
rm(Microamps, Glass, Phosphor)
```

---

URLaddress

*Megabytes Downloaded*

---

### Description

The manager of a URL commercial address is interested in predicting the number of megabytes downloaded, `megasd`, by clients according to the number minutes they are connected, `mconnected`. The manager randomly selects (megabyte, minute) pairs, and records the data. The pairs (`megasd`, `mconnected`) are stored in the data frame `URLaddress`.

### Usage

`URLaddress`

### Format

A data frame with 30 observations on the following 2 variables:

`megasd` megabytes downloaded

`mconnected` number of minutes connected

### Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
xyplot(mconnected~megasd, data = URLaddress, type=c("p","r"))
```

---

vit2005

*Apartments in Vitoria*

---

### Description

Descriptive information and the appraised total price (in Euros) for apartments in Vitoria, Spain.

### Usage

`vit2005`

**Format**

A data frame with 218 observations on the following 16 variables:

`row.labels` the number of the observation

`totalprice` the market total price (in Euros) of the apartment including garage(s) and storage room(s)

`area` the total living area of the apartment in square meters

`zone` a factor indicating the neighborhood where the apartment is located with levels Z11, Z21, Z31, Z32, Z34, Z35, Z36, Z37, Z38, Z41, Z42, Z43, Z44, Z45, Z46, Z47, Z48, Z49, Z52, Z53, Z56, Z61, and Z62.

`category` a factor indicating the condition of the apartment with levels 2A, 2B, 3A, 3B, 4A, 4B, and 5A. The factors are ordered so that 2A is the best and 5A is the worst.

`age` age of the aptment

`floor` floor on which the apartment is located

`rooms` total number of rooms including bedrooms, dining room, and kitchen

`out` a factor indicating the percent of the apartment exposed to the elements. The levels E100, E75, E50, and E25, correspond to complete exposure, 75% exposure, 50% exposure, and 25% exposure respectively.

`conservation` is an ordered factor indicating the state of conservation of the apartment. The levels 1A, 2A, 2B, and 3A are ordered from best to worst conservation.

`toilets` the number of bathrooms

`garage` the number of garages

`elevator` indicates the absence (0) or presence (1) of elevators.

`streetcategory` an ordered factor from best to worst indicating the category of the street with levels S2, S3, S4, and S5

`heating` a factor indicating the type of heating with levels 1A, 3A, 3B, and 4A which correspond to: no heating, low-standard private heating, high-standard private heating, and central heating respectively.

`tras` the number of storage rooms outside of the apartment

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
modTotal <- lm(totalprice~area+as.factor(elevator) +
area:as.factor(elevator), data = vit2005)
modSimpl <- lm(totalprice~area, data = vit2005)
anova(modSimpl,modTotal)
rm(modSimpl, modTotal)
```

---

Wait	<i>Waiting Time</i>
------	---------------------

---

**Description**

A statistician records how long he must wait for his bus each morning. Data are used in Example 10.4.

**Usage**

Wait

**Format**

A data frame with 15 observations on the following variable:

wt waiting time in minutes

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data= Wait,  
wilcox.test(wt,mu=6, alternative="less"))
```

---

Washer	<i>Washer Diameter</i>
--------	------------------------

---

**Description**

Diameter of washers.

**Usage**

Washer

**Format**

A data frame with 20 observations on the following variable:

diameters diameter of washer in cm

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Washer,  
      EDA(diameters))
```

---

Water

*Sodium Content of Water*

---

**Description**

An independent agency measures the sodium content in 20 samples from source X and in 10 samples from source Y and stores them in data frame `Water`. Data are used in Example 9.12.

**Usage**

```
Water
```

**Format**

A data frame with 30 observations on the following 4 variables:

X sodium content measured in mg/L

Y sodium content measured in mg/L

Sodium combined sodium content measured in mg/L

Source a factor with levels X and Y

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Water,  
      t.test(X, Y, alternative="less"))
```

---

WCST

*Wisconsin Card Sorting Test*

---

### Description

The following data are the test scores from a group of 50 patients from the *Virgen del Camino* Hospital (Pamplona, Spain) on the Wisconsin Card Sorting Test.

### Usage

WCST

### Format

A data frame with 50 observations on the following variable:

score score on the Wisconsin Card Sorting Test

### Details

The “Wisconsin Card Sorting Test” is widely used by psychiatrists, neurologists, and neuropsychologists with patients who have a brain injury, neurodegenerative disease, or a mental illness such as schizophrenia. Patients with any sort of frontal lobe lesion generally do poorly on the test.

### Source

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

### Examples

```
densityplot(~score, data = WCST, ref=TRUE)
```

---

WeightGain

*Weight Gain in Rats*

---

### Description

The data come from an experiment to study the gain in weight of rats fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal).

### Usage

```
data(WeightGain)
```

**Format**

A data frame with 40 observations on the following 4 variables.

ProteinSource a factor with levels Beef Cereal

ProteinAmount a factor with levels High Low

weightgain weight gain in grams

**Details**

The design of the experiment is a completely randomized with ten rats on each of the four treatments.

**Source**

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994) *A Handbook of Small Datasets*. Chapman and Hall/CRC, London.

**Examples**

```
aov(weightgain~ProteinSource*ProteinAmount, data=WeightGain)
```

---

WheatSpain

*Wheat Surface Area in Spain*

---

**Description**

Seventeen Spanish communities and their corresponding surface area (in hectares) dedicated to growing wheat

**Usage**

WheatSpain

**Format**

A data frame with 17 observations on the following 3 variables:

community a factor with levels Andalucia, Aragon, Asturias, Baleares, C.Valenciana, Canarias, Cantabria, Castilla-La Mancha, Castilla-Leon, Cataluna, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarra, and P.Vasco

hectares surface area measured in hectares

acres surface area measured in acres

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = WheatSpain,  
      boxplot(hectares))
```

---

wheatUSA2004	<i>USA Wheat Surface 2004</i>
--------------	-------------------------------

---

**Description**

USA's 2004 harvested wheat surface by state

**Usage**

```
wheatUSA2004
```

**Format**

A data frame with 30 observations on the following 2 variables.

STATES a factor with levels AR, CA, CO, DE, GA, ID, IL, IN, KS, KY, MD, MI, MO, MS, MT, NC, NE, NY, OH, OK, OR, Other, PA, SC, SD, TN, TX, VA, WA, and WI

ACRES wheat surface area measured in 1000s of acres

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = wheatUSA2004,  
      hist(ACRES))
```

---

wilcoxE.test	<i>Wilcox Exact Test</i>
--------------	--------------------------

---

**Description**

Performs exact one sample and two sample Wilcoxon tests on vectors of data.

**Usage**

```
wilcoxE.test(x, y = NULL, mu = 0, paired = FALSE,  
            alternative = "two.sided", conf.level = 0.95)
```

**Arguments**

x	is a numeric vector of data values. Non-finite (i.e. infinite or missing) values will be omitted.
y	an optional numeric vector of data values
mu	a number specifying an optional parameter used to form the null hypothesis
paired	a logical indicating whether you want a paired test
alternative	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater", or "less". The initial letter only may be given.
conf.level	confidence level of the interval

**Details**

If only x is given, or if both x and y are given and paired="TRUE", a Wilcoxon signed rank test of the null hypothesis that the distribution of x (in the one sample case) or of  $x - y$  (in the paired two sample case) is symmetric about mu is performed.

Otherwise, if both x and y are given and paired="FALSE", a Wilcoxon rank sum test is done. In this case, the null hypothesis is that the distributions of x and y differ by a location shift of mu and the alternative is that they differ by some other location shift (and the one-sided alternative "greater" is that x is shifted to the right of y).

**Value**

A list with class htest containing the following components:

statistic	the value of the test statistic with a name describing it
p.value	the p-value for the test
null.value	the location parameter mu
alternative	a character string describing the alternative hypothesis
method	the type of test applied
data.name	a character string giving the names of the data
conf.int	a confidence interval for the location parameter
estimate	an estimate of the location parameter

**Note**

The function is rather primitive and should only be used for problems with fewer than 19 observations as the memory requirements are rather large.

**Author(s)**

Alan T. Arnholt

**References**

Gibbons, J.D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference*. Marcel Dekker Inc., New York.

Myles Hollander & Douglas A. Wolfe (1999), *Nonparametric Statistical Inference*. New York: John Wiley & Sons.

**See Also**

[wilcox.test](#)

**Examples**

```
# Wilcoxon Signed Rank Test - Example 10.3
PH <- c(7.2,7.3,7.3,7.4)
wilcoxE.test(PH, mu=7.25, alternative="greater")
# Wilcoxon Signed Rank Test (Dependent Samples) - Example 10.5 part c.
with(data = Aggression,
wilcoxE.test(violence, noviolence, paired=TRUE, alternative="greater"))
# Wilcoxon Rank Sum Test - Example 10.7
x <- c(7.2,7.2,7.3,7.3)
y <- c(7.3,7.3,7.4,7.4)
wilcoxE.test(x,y)
rm(PH, x, y)
```

---

Wool

*Wool Production*

---

**Description**

Random sample of wool production in kilograms on 5 different days at two different locations

**Usage**

Wool

**Format**

A data frame with 15 observations on the following 2 variables:

textileA wool production in thousands of kilograms

textileB wool production in thousands of kilograms

**Source**

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2008) *Probability and Statistics with R*. Chapman & Hall/CRC.

**Examples**

```
with(data = Wool,
      t.test(textileA, textileB))
```

---

z.test

*Z-test*


---

**Description**

This function is based on the standard normal distribution and creates confidence intervals and tests hypotheses for both one and two sample problems.

**Usage**

```
z.test(x, y = NULL, alternative = "two.sided", mu = 0, sigma.x = NULL,
       sigma.y = NULL, conf.level = 0.95)
```

**Arguments**

x	numeric vector; NAs and Infs are allowed but will be removed.
y	numeric vector; NAs and Infs are allowed but will be removed.
alternative	character string, one of "greater", "less" or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, alternative refers to the true mean of the parent population in relation to the hypothesized value mu. For the standard two-sample tests, alternative refers to the difference between the true population mean for x and that for y, in relation to mu.
mu	a single number representing the value of the mean or difference in means specified by the null hypothesis
sigma.x	a single number representing the population standard deviation for x
sigma.y	a single number representing the population standard deviation for y
conf.level	confidence level for the returned confidence interval, restricted to lie between zero and one

**Details**

If y is NULL, a one-sample z-test is carried out with x. If y is not NULL, a standard two-sample z-test is performed.

**Value**

A list of class `htest`, containing the following components:

<code>statistic</code>	the z-statistic, with names attribute "z"
<code>p.value</code>	the p-value for the test
<code>conf.int</code>	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute <code>conf.level</code> . When <code>alternative</code> is not "two.sided", the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values $k$ for which one would not reject the null hypothesis that the true mean or difference in means is $k$ . Here infinity will be represented by <code>Inf</code> .
<code>estimate</code>	vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	is the value of the mean or difference in means specified by the null hypothesis. This equals the input argument <code>mu</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : "greater", "less" or "two.sided".
<code>data.name</code>	a character string (vector of length 1) containing the actual names of the input vectors <code>x</code> and <code>y</code>

**Null Hypothesis**

For the one-sample z-test, the null hypothesis is that the mean of the population from which `x` is drawn is  $\mu$ . For the standard two-sample z-tests, the null hypothesis is that the population mean for `x` less than for `y` is  $\mu$ .

The alternative hypothesis in each case indicates the direction of divergence of the population mean for `x` (or difference of means for `x` and `y`) from  $\mu$  (i.e., "greater", "less", "two.sided").

**Test Assumptions**

The assumption of normality for the underlying distribution or a sufficiently large sample size is required along with the population standard deviation to use Z procedures.

**Confidence Interval**

For each of the above tests, an expression for the related confidence interval (returned component `conf.int`) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when `alternative` is not "two.sided"; infinity will be represented by `Inf`.

**Author(s)**

Alan T. Arnholt

## References

Kitchens, L.J. (2003). *Basic Statistics and Data Analysis*. Duxbury.

Hogg, R. V. and Craig, A. T. (1970). *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

## See Also

[zsum.test](#), [tsum.test](#)

## Examples

```
with(data = Grocery,
z.test(x=groceries,sigma.x=30,conf.level=.97)$conf)
# Example 8.3 from PASWR.

x <- rnorm(12)
z.test(x,sigma.x=1)
# Two-sided one-sample z-test where the assumed value for
# sigma.x is one. The null hypothesis is that the population
# mean for 'x' is zero. The alternative hypothesis states
# that it is either greater or less than zero. A confidence
# interval for the population mean will be computed.

x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7., 6.4, 7.1, 6.7, 7.6, 6.8)
y <- c(4.5, 5.4, 6.1, 6.1, 5.4, 5., 4.1, 5.5)
z.test(x, sigma.x=0.5, y, sigma.y=0.5, mu=2)
# Two-sided standard two-sample z-test where both sigma.x
# and sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is 2.
# The alternative hypothesis is that this difference is not 2.
# A confidence interval for the true difference will be computed.

z.test(x, sigma.x=0.5, y, sigma.y=0.5, conf.level=0.90)
# Two-sided standard two-sample z-test where both sigma.x and
# sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is zero.
# The alternative hypothesis is that this difference is not
# zero. A 90% confidence interval for the true difference will
# be computed.

rm(x, y)
```

---

zsum.test

*Summarized z-test*


---

### Description

This function is based on the standard normal distribution and creates confidence intervals and tests hypotheses for both one and two sample problems based on summarized information the user passes to the function. Output is identical to that produced with `z.test`.

### Usage

```
zsum.test(mean.x, sigma.x = NULL, n.x = NULL, mean.y = NULL,
sigma.y = NULL, n.y = NULL, alternative = "two.sided", mu = 0,
conf.level = 0.95)
```

### Arguments

<code>mean.x</code>	a single number representing the sample mean of x
<code>sigma.x</code>	a single number representing the population standard deviation for x
<code>n.x</code>	a single number representing the sample size for x
<code>mean.y</code>	a single number representing the sample mean of y
<code>sigma.y</code>	a single number representing the population standard deviation for y
<code>n.y</code>	a single number representing the sample size for y
<code>alternative</code>	is a character string, one of "greater", "less" or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, <code>alternative</code> refers to the true mean of the parent population in relation to the hypothesized value <code>mu</code> . For the standard two-sample tests, <code>alternative</code> refers to the difference between the true population mean for x and that for y, in relation to <code>mu</code> .
<code>mu</code>	a single number representing the value of the mean or difference in means specified by the null hypothesis
<code>conf.level</code>	confidence level for the returned confidence interval, restricted to lie between zero and one

### Details

If `y` is `NULL`, a one-sample z-test is carried out with `x`. If `y` is not `NULL`, a standard two-sample z-test is performed.

### Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the z-statistic, with names attribute <code>z</code> .
<code>p.value</code>	the p-value for the test

conf.int	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute conf.level. When alternative is not "two.sided", the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values $k$ for which one would not reject the null hypothesis that the true mean or difference in means is $k$ . Here, infinity will be represented by Inf.
estimate	vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component estimate has a names attribute describing its elements.
null.value	the value of the mean or difference in means specified by the null hypothesis. This equals the input argument mu. Component null.value has a names attribute describing its elements.
alternative	records the value of the input argument alternative: "greater" , "less" or "two.sided".
data.name	a character string (vector of length 1) containing the names $x$ and $y$ for the two summarized samples

### Null Hypothesis

For the one-sample z-test, the null hypothesis is that the mean of the population from which  $x$  is drawn is  $\mu$ . For the standard two-sample z-tests, the null hypothesis is that the population mean for  $x$  less that for  $y$  is  $\mu$ .

The alternative hypothesis in each case indicates the direction of divergence of the population mean for  $x$  (or difference of means of  $x$  and  $y$ ) from  $\mu$  (i.e., "greater" , "less" , "two.sided" ).

### Test Assumptions

The assumption of normality for the underlying distribution or a sufficiently large sample size is required along with the population standard deviation to use Z procedures.

### Confidence Intervals

For each of the above tests, an expression for the related confidence interval (returned component conf.int) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of conf.int, the confidence interval will be half-infinite when alternative is not "two.sided"; infinity will be represented by Inf.

### Author(s)

Alan T. Arnholt

### References

- Kitchens, L. J. (2003). *Basic Statistics and Data Analysis*. Duxbury.
- Hogg, R. V. and Craig, A. T. (1970). *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.
- Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

### See Also

[z.test](#), [tsum.test](#)

### Examples

```
zsum.test(mean.x=56/30,sigma.x=2, n.x=30, alternative="greater", mu=1.8)
# Example 9.7 part a. from PASWR.
x <- rnorm(12)
zsum.test(mean(x),sigma.x=1,n.x=12)
# Two-sided one-sample z-test where the assumed value for
# sigma.x is one. The null hypothesis is that the population
# mean for 'x' is zero. The alternative hypothesis states
# that it is either greater or less than zero. A confidence
# interval for the population mean will be computed.
# Note: returns same answer as:
z.test(x,sigma.x=1)
#
x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)
y <- c(4.5, 5.4, 6.1, 6.1, 5.4, 5.0, 4.1, 5.5)
zsum.test(mean(x), sigma.x=0.5, n.x=11 ,mean(y), sigma.y=0.5, n.y=8, mu=2)
# Two-sided standard two-sample z-test where both sigma.x
# and sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is 2.
# The alternative hypothesis is that this difference is not 2.
# A confidence interval for the true difference will be computed.
# Note: returns same answer as:
z.test(x, sigma.x=0.5, y, sigma.y=0.5)
#
zsum.test(mean(x), sigma.x=0.5, n.x=11, mean(y), sigma.y=0.5, n.y=8,
conf.level=0.90)
# Two-sided standard two-sample z-test where both sigma.x and
# sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is zero.
# The alternative hypothesis is that this difference is not
# zero. A 90% confidence interval for the true difference will
# be computed. Note: returns same answer as:
z.test(x, sigma.x=0.5, y, sigma.y=0.5, conf.level=0.90)
rm(x, y)
```

# Index

## \*Topic **aplot**

oneway.plots, 43  
twoway.plots, 71

## \*Topic **arith**

Combinations, 17  
SRS, 57

## \*Topic **datasets**

Aggression, 4  
Apple, 5  
AptSize, 6  
Baberuth, 7  
Bac, 8  
Battery, 9  
biomass, 10  
Bodyfat, 11  
Calculus, 12  
Cars2004EU, 12  
Chips, 14  
CircuitDesigns, 15  
CosAma, 18  
Cows, 18  
Depend, 19  
Drosophila, 20  
Engineer, 21  
EPIDURAL, 22  
EPIDURALf, 23  
EURD, 24  
fagus, 24  
FCD, 25  
Fertilize, 26  
food, 27  
Formula1, 27  
GD, 28  
glucose, 29  
Grades, 29  
Grocery, 30  
HardWater, 31  
House, 32  
HSwrestler, 32

Hubble, 33  
InsurQuotes, 34  
janka, 35  
Kinder, 36  
Leddiode, 38  
LostR, 39  
MilkCarton, 39  
Phenyl, 43  
Phone, 44  
Rat, 45  
Ratbp, 46  
Refrigerator, 46  
Roacheggs, 47  
Salinity, 48  
satfruit, 49  
SBIQ, 50  
Schizo, 51  
Score, 51  
SDS4, 52  
SimDataST, 55  
SimDataXT, 56  
Soccer, 56  
StatTemps, 58  
Stschool, 59  
Sundig, 59  
sunflower, 60  
SurfaceSpain, 61  
Swintimes, 62  
Tennis, 63  
TestScores, 63  
Tire, 64  
TireWear, 65  
titanic3, 65  
TOE, 67  
Top20, 67  
URLaddress, 72  
vit2005, 72  
Wait, 74  
Washer, 74

- Water, 75
- WCST, 76
- WeightGain, 76
- WheatSpain, 77
- wheatUSA2004, 78
- Wool, 80
- \*Topic **distribution**
  - bino.gen, 9
  - CIsim, 15
  - ksdist, 37
  - ksLdist, 37
  - normarea, 40
  - ntester, 42
- \*Topic **htest**
  - SIGN.test, 53
  - tsum.test, 68
  - wilcoxE.test, 78
  - z.test, 81
  - zsum.test, 84
- \*Topic **misc**
  - interval.plot, 35
- \*Topic **package**
  - PASWR-package, 4
- \*Topic **regression**
  - checking.plots, 13
- \*Topic **univar**
  - EDA, 20
  - nsize, 41
- Aggression, 4
- Apple, 5
- AptSize, 6
- Baberuth, 7
- Bac, 8
- Battery, 9
- bino.gen, 9
- biomass, 10
- Bodyfat, 11
- Calculus, 12
- Cars2004EU, 12
- checking.plots, 13
- Chips, 14
- CircuitDesigns, 15
- CIsim, 15, 35
- Combinations, 17, 58
- CosAma, 18
- Cows, 18
- Depend, 19
- Drosophila, 20
- EDA, 20
- Engineer, 21
- EPIDURAL, 22
- EPIDURALf, 23
- EURD, 24
- fagus, 24
- FCD, 25
- Fertilize, 26
- food, 27
- Formula1, 27
- GD, 28
- glucose, 29
- Grades, 29
- Grocery, 30
- HardWater, 31
- House, 32
- HSwrestler, 32
- Hubble, 33
- InsurQuotes, 34
- interval.plot, 35
- janka, 35
- Kinder, 36
- ksdist, 37, 38
- ksLdist, 37, 37
- Leddiode, 38
- LostR, 39
- MilkCarton, 39
- normarea, 40
- nsize, 41
- ntester, 42
- oneway.plots, 43, 71
- PASWR (PASWR-package), 4
- PASWR-package, 4
- Phenyl, 43
- Phone, 44
- Rat, 45

Ratbp, 46  
Refrigerator, 46  
Roacheggs, 47

Salinity, 48  
satfruit, 49  
SBIQ, 50  
Schizo, 51  
Score, 51  
SDS4, 52  
SIGN.test, 53  
SimDataST, 55  
SimDataXT, 56  
Soccer, 56  
SRS, 17, 57  
StatTemps, 58  
Stschool, 59  
Sundig, 59  
sunflower, 60  
SurfaceSpain, 61  
Swintimes, 62

Tennis, 63  
TestScores, 63  
Tire, 64  
TireWear, 65  
titanic3, 65  
TOE, 67  
Top20, 67  
tsum.test, 54, 68, 83, 86  
twoway.plots, 14, 43, 71

URLaddress, 72

vit2005, 72

Wait, 74  
Washer, 74  
Water, 75  
WCST, 76  
WeightGain, 76  
WheatSpain, 77  
wheatUSA2004, 78  
wilcox.test, 80  
wilcoxE.test, 78  
Wool, 80

z.test, 54, 70, 81, 86  
zsum.test, 54, 70, 83, 84